Recovery and convergence rate of the Frank-Wolfe Algorithm for the m-EXACT-SPARSE Problem

Farah Cherfaoui, Valentin Emiya, Liva Ralaivola, Sandrine Anthoine

Recovery and convergence rate of the Frank-Wolfe Algorithm for the m-EXACT-SPARSE Problem

Farah Cherfaoui, Valentin Emiya, Liva Ralaivola and Sandrine Anthoine

***Abstract*—We study the properties of the Frank-Wolfe algorithm to solve the m-EXACT-SPARSE reconstruction problem, where a signal y must be expressed as a sparse linear combination of a predefined set of atoms, called *dictionary*. We prove that when the dictionary is *quasi-incoherent*, then the iterative process implemented by the Frank-Wolfe algorithm only recruits atoms from the support of the signal, that is the smallest set of atoms from the dictionary that allows for a perfect reconstruction of y. We also prove that when the dictionary is quasi-incoherent, there exists an iteration beyond which the algorithm converges exponentially.**

***Index Terms*—sparse representation, Frank-Wolfe algorithm, recovery properties, exponential convergence.**

I. INTRODUCTION

A. The m-EXACT-SPARSE problem

A signal y in $\mathbb{R}^d$ is m-sparse in a given dictionary Φ in $\mathbb{R}^{d\times n}$, when y is a linear combination of at most m atoms (i.e. columns) of Φ. Given a dictionary Φ and an m-sparse signal y, the m-EXACT-SPARSE problem is to express y as a linear expansion of at most m atoms.

Here, we study the algorithmic properties of the Frank-Wolfe optimization procedure [5] when implemented to solve this problem.

In the sequel, the dictionary $\Phi = [\varphi_1 \cdots \varphi_n] \in \mathbb{R}^{d\times n}$ is the matrix made of the n atoms $\varphi_1, \ldots, \varphi_n \in \mathbb{R}^d$, assumed to be so that $\|\varphi_i\|_2 = 1, \forall i$. The support Λ of an m-sparse signal y is the smallest subset of $\{1, \ldots, n\}$ of size m such that y is in the span of the atoms indexed by Λ (we give in Section III the condition under which this support is unique). The sparsity of a linear combination Φx ($x \in \mathbb{R}^n$) such that $y = \Phi x$ is measured by the number of nonzero entries of x, sometimes referred to as the quasi-norm $\|x\|_0$ of x.

Formally the m-EXACT-SPARSE problem is the following. Giving a dictionary Φ and an m-sparse signal y:

$$\text{find } x \text{ s.t. } y = \Phi x \text{ and } \|x\|_0 \leq m.$$

Since y is a linear combination of at most m atoms of Φ, a solution of the problem can be obtained by solving:

$$\arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|y - \Phi x\|_2^2 \;\text{ s.t. } \|x\|_0 \leq m. \qquad (1)$$

F. Cherfaoui, V. Emiya are with Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France.

L. Ralaivola is with Aix Marseille Univ, Université de Toulon, CNRS, IUF and Criteo.

S. Anthoine is with Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France.

B. Related work

The m-EXACT-SPARSE problem is NP-hard [4]. This raises the question on the conditions under which m-EXACT-SPARSE becomes tractable. Many types of algorithms have been proposed to answer this question, including brute force [11], nonlinear programming [13], and greedy pursuits [10], [12], [15], [14], [1]. Another way to tackle the m-EXACT-SPARSE problem is to relax Problem (1) for example, by changing the l_0 quasi-norm to an l_1 norm (e.g. LASSO [16] and Basis Pursuit [3]). This gives rise to convex optimization problems for which a number of different methods can be used. Moreover the approximation that is made by such relaxation can sometimes be controlled. Even better, Tropp in [18] proved that under specific conditions, the supports of the solutions of the l_1-relaxation of LASSO and Basis Pursuit are included in that of the m-EXACT-SPARSE problem.

In this paper, we consider the following l_1-relaxation:

$$\arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|y - \Phi x\|_2^2 \;\text{ s.t. } \|x\|_1 \leq \beta \qquad (2)$$

where $\|.\|_1$ is the l_1 norm and $\beta > 0$ is a hyperparameter. We will solve this problem by applying a Frank-Wolfe procedure [5].

To motivate our work, let us go back to the m-EXACT-SPARSE problem. Matching Pursuit (MP) [10] and Orthogonal Matching Pursuit (OMP) [12] are two greedy algorithms used to seek a solution to this problem. These two algorithms were studied a lot, and recovery guarantees for sparse signals were given for MP and OMP. Tropp [17] and Gribonval and Vandergheynst [6] proved that, if the dictionary is quasi-incoherent (for example, the dictionary is close to an orthonormal basis), then at each iteration the MP and OMP algorithms pick up an atom indexed by the support. They also proved that MP converges exponentially, and that OMP converges after exactly m iterations, where m is the size of the support. In their papers, Tropp [17] and Gribonval and Vandergheynst [6] also studied the case when the signal is not exactly m-sparse, and proved similar results.

On the other hand, the Frank-Wolfe algorithm [5] is an iterative optimization algorithm designed for constrained convex optimization. It has been proven to converge exponentially if the objective function is strongly convex [7] and linearly in the other cases [5]. When solving (2), the atom selection steps in Matching Pursuit and Frank-Wolfe are very similar (as we will see in Section II-C). This inspired for example Jaggi and al. [8] to exploit tools usually used for the Frank-Wolfe algorithm to prove the convergence of the MP algorithm when no assumptions are made on the dictionary.

In our paper, we will do the opposite. We will exploit tools used for MP and OMP to study the properties of the Frank-Wolfe algorithm when seeking a solution of Problem (2) when y is m-sparse.

C. Main results

We show that the Frank Wolfe algorithm, when used to solve (2), enjoys recovery and convergence properties regarding m-EXACT-SPARSE that are similar to those established in [6], [17] for MP and OMP, under the very same assumptions.

Our results rely on a fundamental quantity associated to a dictionary $\Phi = [\varphi_1 \cdots \varphi_n]$: its Babel function, defined as

$$\mu_1(m) = \max_{|\Lambda|=m} \max_{i \notin \Lambda} \sum_{j \in \Lambda} |\langle \varphi_i, \varphi_j \rangle|.$$

For given m, $\mu_1(m)$ is roughly a measure on how well any atom from Φ can be expressed as a linear combination of a set of $m-1$ other atoms. When $m = 1$, the Babel function boils down to

$$\mu = \mu_1(1) = \max_{j \neq k} |\langle \varphi_j, \varphi_k \rangle|,$$

which is known as the *coherence* of Φ. When

$$m < \tfrac{1}{2}(\mu^{-1} + 1),$$

we say that the dictionary is *quasi-incoherent*. In Featured Theorem 1, we state that when the dictionary is quasi-incoherent, then at each iteration the Frank-Wolfe algorithm picks up an atom indexed by the support of y.

Featured Theorem 1. *Let Φ be a quasi-incoherent dictionary, and y an m-sparse signal. Then at each iteration, the Frank-Wolfe algorithm picks up an atom of the support of the signal.*

For such dictionaries, we also prove that the rate of convergence of the Frank-Wolfe algorithm is exponential beyond a certain iteration even though the function we consider is not strongly convex. This is given by Featured Theorem 2.

Featured Theorem 2. *Let Φ be a quasi-incoherent dictionary and y an m-sparse signal. Under some conditions on y and β, there exists an iteration K of the Frank-Wolfe algorithm and $0 < \theta \leq 1$ such that:*

$$\|y - \Phi x_{k+1}\|_2^2 \leq \|y - \Phi x_k\|_2^2 (1 - \theta) \qquad \forall\ k \geq K$$

where θ depends on $\mu_1(m-1)$, β, and y and $0 < \theta \leq 1$ (which implies the exponential convergence).

D. Organization of the Paper

In Section II, we instantiate the Frank-Wolfe algorithm for Problem (2) and we relate it to MP and OMP. Section III is devoted to the statement and the proofs of our main results. We probe their optimality with numerical experiments in Section IV.

II. THE ALGORITHMS

This section recalls Matching Pursuit and Orthogonal Matching Pursuit, the classical greedy algorithms used for solving the m-EXACT-SPARSE problem. We then present the Frank-Wolfe algorithm and derive it for Problem (2), showing its similarities with MP and OMP.

A. Matching Pursuit and Orthogonal Matching Pursuit

Let Φ be a dictionary and y an exactly m-sparse signal (i.e. there exists an x with exactly m non-zero entries such that $y = \Phi x$). If the dictionary is an orthonormal basis, the m-EXACT-SPARSE problem has an easy solution: one chooses the m atoms having the largest absolute inner products with the signal. These m atoms have non-zero inner products with the signal. The corresponding non-zero values in x are then the inner products themselves.

Algorithmically, this can be achieved by building the approximation one term at a time. Noting y_k the current approximation and $r_k = y - y_k$ the corresponding residual, we select at each time the atom which has the largest inner product with the residual, and update the approximation.

Matching Pursuit [10] and Orthogonal Matching Pursuit [12] are two greedy algorithms used for approximating signals that build upon this idea of greedy selection and iterative updates. MP and OMP initialize the first approximation $y_0 = 0$ and residual $r_0 = y$ and then repeat the following steps:

- Atom selection: $\lambda_k = \arg\max_i |\langle \varphi_i, r_k \rangle|$.
- Approximation update:
  - MP: $y_{k+1} = y_k + \langle \varphi_{\lambda_k}, r_k \rangle \varphi_{\lambda_k}$
  - OMP: $y_{k+1} = \arg\min_{a \in \mathrm{span}(\{\varphi_{\lambda_0}, \ldots, \varphi_{\lambda_k}\})} \|y - a\|_2$
- Residual update: $r_{k+1} = y - y_{k+1}$.

B. Frank-Wolfe

The Frank-Wolfe algorithm [5] is an iterative algorithm developed to solve the optimization problem:

$$\min_{x \in \mathcal{C}} f(x) \quad \text{s.t. } x \in \mathcal{C} \tag{3}$$

where f is a convex and continuously differentiable function and $\mathcal{C}$ is a compact and convex set. The Frank-Wolfe algorithm is initialized with an element of $\mathcal{C}$. Then, at iteration k, the algorithm applies the three following steps:

- Descent direction selection: $s_k = \arg\min_{s \in \mathcal{C}} \langle s, \nabla f(x_k) \rangle$.
- Step size optimization: $\gamma_k = \arg\min_{\gamma \in [0,1]} f((1-\gamma)x_k + \gamma s_k)$.
- Update: $x_{k+1} = (1 - \gamma_k)x_k + \gamma_k s_k$.

Note that the step-size γ_k can be chosen by others methods [8], without affecting the convergence properties of the algorithm.

C. Frank-Wolfe for the m-EXACT-SPARSE problem

We are interested in solving the m-EXACT-SPARSE problem by finding the solution of the following problem:

$$\arg\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|y - \Phi x\|_2^2 \quad \text{s.t. } \|x\|_1 \leq \beta \tag{2}$$

Algorithm 1: The Frank-Wolfe algorithm

Data: signal y, dictionary $\Phi = [\varphi_1, \dots, \varphi_n], \beta > 0$.

$x_0 = 0$
$k = 0$
while *stopping criterion not verified* **do**
  $i_k = \arg\max_{i \in \{1,\dots,n\}} |\langle \varphi_i, y - \Phi x_k \rangle|$
  $s_k = \text{sign}(\langle \varphi_{i_k}, y - \Phi x_k \rangle) \beta e_{i_k}$
  $\gamma_k = \arg\min_{\gamma \in [0,1]} \|y - \Phi(x_k + \gamma(s_k - x_k))\|_2^2$
  $x_{k+1} = x_k + \gamma_k(s_k - x_k)$
  $k = k + 1$
end

using the Frank-Wolfe algorithm. Let us derive the algorithm in this context. The objective function to minimize is:

$$f(x) = \tfrac{1}{2}\|y - \Phi x\|_2^2,$$

and the convex set is:

$$\mathcal{C} = \{x : \ \|x\|_1 \leq \beta\}.$$

$\mathcal{C}$ is the l_1 ball of radius β that we denote by $\mathcal{B}_1(\beta)$. It can be written as the convex hull: $\mathcal{B}_1(\beta) = \text{conv}\{\pm\beta e_i | i \in \{1, \dots, n\}\}$, with e_i the canonical basis vectors of $\mathbb{R}^n$. Moreover, $\nabla f(x) = \Phi^t(\Phi x - y)$. The selection step of the Frank-Wolfe algorithm thus becomes:

$$s_k = \arg\min_{s \in \text{conv}\{\pm\beta e_i | i \in \{1,\dots,n\}\}} \langle s, \Phi^t(\Phi x_k - y) \rangle.$$

Since this optimization problem is linear and $\mathcal{B}_1(\beta)$ is closed and bounded, the solution corresponds to an extreme point of $\mathcal{B}_1(\beta)$ (see [2] for more details):

$$s_k = \arg\min_{s \in \{\pm\beta e_i | i \in \{1,\dots,n\}\}} \langle s, \Phi^t(\Phi x_k - y) \rangle$$

or equivalently

$$s_k = \arg\max_{s \in \{\pm\beta e_i | i \in \{1,\dots,n\}\}} \langle \Phi s, y - \Phi x_k \rangle.$$

Noticing that $s = \pm\beta e_i$ implies $\Phi s = \pm\beta\varphi_i$, we conclude that the direction selection step of the Frank-Wolfe algorithm for Problem (2) can be rewritten as:

$$\begin{cases} i_k & = \arg\max_{i \in \{1,\dots,n\}} |\langle \varphi_i, y - \Phi x_k \rangle| \\ s_k & = \text{sign}(\langle \varphi_{i_k}, y - \Phi x_k \rangle) \beta e_{i_k}. \end{cases}$$

Recalling that the residual r_k is:

$$r_k = y - \Phi x_k,$$

we notice that we have the same atom selection as in MP and OMP: $i_k = \lambda_k = \arg\max_i |\langle \varphi_i, r_k \rangle|$. Finally, we specify the initialization $x_0 = 0$ which is in $\mathcal{B}_1(\beta)$. This completes the description of the Frank-Wolfe algorithm for Problem (2) which is summarized in Algorithm 1.

In the sequel, we will be interested in the recovery and convergence properties of this algorithm in the case when the dictionary is *quasi-incoherent* ($m < \frac{1}{2}(\mu^{-1} + 1)$). This hypothesis implies that the atoms of any subset of at most m atoms ($\{\varphi_i | i \in \Lambda\}$ such that $|\Lambda| \leq m$) are necessarily linearly independent and also that for any m-sparse signal y, the expansion coefficients x^* such that $y = \Phi x^*$ and $\|x^*\|_0 \leq m$ and the corresponding support are unique.

In this case, x^* is the unique solution of the m-EXACT-SPARSE problem and also of Problem (1). As for the relaxed problem (2), we have $f(x^*) = 0$; however x^* is a solution of Problem (2) if and only if it is feasible i.e. $\|x^*\|_1 \leq \beta$.

Now, let us clarify some notations. For an m-sparse signal $y = \Phi x^*$, we denote by Λ_{opt} its support i.e. $y = \sum_{i \in \Lambda_{opt}} x^*[i]\varphi_i$ such that $|\Lambda_{opt}| \leq m$. For Λ a subset of $\{1, \dots, n\}$, we denote by Φ_Λ the matrix whose columns are the atoms indexed by Λ. When Λ is the support Λ_{opt}, we note λ^*_{min} (resp. λ^*_{max}) its lowest (resp. largest singular value). For a matrix Φ we denote by $\text{span}(\Phi)$ the vector space spanned by its columns. Finally, when we study the convergence of Algorithm 1, we consider the residual squared norm as it is linked to the objective function:

$$f(x_k) = \tfrac{1}{2}\|y - \Phi x_k\|_2^2 = \tfrac{1}{2}\|r_k\|_2^2.$$

III. EXACT RECOVERY AND EXPONENTIAL CONVERGENCE

In this section, we state and prove our main results on the recovery property and the convergence rate of Algorithm 1. We state in Theorem 1 the recovery guaranties of this algorithm, and we present its convergence rate in Theorem 2.

A. Recovery condition

Tropp [17] and Gribonval and Vandergheynst [6] proved that when the dictionary is quasi-incoherent, then MP and OMP exactly recover the m-expansion of any m-sparse signal. To do so, they prove that at each step, MP and OMP pick an atom of the support. Theorem 1 extends this result to the Frank-Wolfe algorithm.

Theorem 1. *Let Φ be a dictionary, μ its coherence, and y an m-sparse signal of support Λ_{opt}. If $m < \frac{1}{2}(\mu^{-1} + 1)$, then at each iteration, Algorithm 1 picks up a correct atom, i.e. $\forall\ k$, $i_k \in \Lambda_{opt}$.*

Remark 1. *As in [17], [6], the condition $m < \frac{1}{2}(\mu^{-1} + 1)$ can be replaced by a weaker support-specific condition called the exact recovery condition (ERC): $\max_{i \notin \Lambda_{opt}} \|\Phi^+_{\Lambda_{opt}}\varphi_i\|_1 < 1$. ERC is not easy to check in practice because it depends on the unknown support Λ_{opt}. The quasi-incoherence $m < \frac{1}{2}(\mu^{-1} + 1)$ is a sufficient condition to have ERC [17]. Furthermore, this last condition is easy to check.*

Proof of Theorem 1. The proof of this theorem is very similar to the proof of Theorem 3.1 in [17]. One shows by induction that at each step the residual $r_k = y - \Phi x_k$ remains in $\text{span}(\Phi_{\lambda_{opt}})$ and in the process that the selected atom is in Λ_{opt}.

- $k = 0$: by definition $r_0 = y$ is in $\text{span}(\Phi_{\lambda_{opt}})$.
- If $k \geq 0$: we assume that $r_k \in \text{span}(\Phi_{\Lambda_{opt}})$. The atom φ_{i_k} is a "good atom" (i.e. $i_k \in \Lambda_{opt}$), if and only if:

$$\rho(r_k) = \frac{\|\Psi^t_{opt} r_k\|_\infty}{\|\Phi^t_{opt} r_k\|_\infty} < 1.$$

Tropp [17] proved that $\rho(r_k) \leq \max_{i \notin \Lambda_{opt}} \|\Phi^+_{\Lambda_{opt}} \varphi_i\|_1 \leq \frac{\mu_1(m)}{1-\mu_1(m-1)}$ and that $m < \frac{1}{2}(\mu^{-1}+1)$ implies $\frac{\mu_1(m)}{1-\mu_1(m-1)} < 1$.
Hence i_k is in Λ_{opt} and $s_k = \pm\beta\varphi_{i_k}$ is thus in $\mathrm{span}(\Phi_{\Lambda_{opt}})$. Notice that $r_{k+1} = r_k - \gamma_k \Phi(s_k - x_k)$. Since r_k is also $\mathrm{span}(\Phi_{\Lambda_{opt}})$ by assumption, we deduce that r_{k+1} is in $\mathrm{span}(\Phi_{\Lambda_{opt}})$.

□

Theorem 1 specifies that under the quasi-incoherence hypothesis, Algorithm 1 only recruits atoms of the support of y. As noted before however, the expansion x^* can not always be reached (because it might be the case that $\|x^*\|_1 > \beta$). In the case when it can be reached (i.e. when $\|x^*\|_1 \leq \beta$) one can furthermore prove that the expansion itself is recovered:

Corollary 1. *Let Φ be a dictionary, μ its coherence, and y an m-sparse signal of support Λ_{opt}. If $m < \frac{1}{2}(\mu^{-1}+1)$ and $\|x^*\|_1 \leq \beta$ then the sequence x_k converges to x^* (i.e. Algorithm 1 exactly recovers the m-sparse expansion of y).*

Proof of Corollary 1. $\|x^*\|_1 \leq \beta$ so that x^* is a solution of Problem (2). The Frank-Wolfe algorithm is known to converge in terms of objective values ($f(x_k)$), we deduce that $f(x_k)$ converges to $f(x^*) = 0$. Since Theorem 1 ensures that the iterates x_k are in $\mathrm{span}(\Phi_{\Lambda_{opt}})$, we also have convergence of the iterates (x_k converge to x^*) since

$$\begin{aligned} |f(x_k) - f(x^*)| = |f(x_k) - 0| &= \tfrac{1}{2}\|y - \Phi x_k\|_2^2 \\ &= \tfrac{1}{2}\|\Phi x^* - \Phi x_k\|_2^2 \\ &\geq \tfrac{1}{2}\left(\lambda^*_{min}\right)^2 \|x^* - x_k\|_2^2 \end{aligned}$$

where the last line holds because $x_k - x^*$ is in $\mathrm{span}(\Phi_{\Lambda_{opt}})$ and $\lambda^*_{min} > 0$ since $m < \frac{1}{2}(\mu^{-1}+1)$ [17]. □

In this section, we presented the recovery guarantees for the Frank-Wolfe algorithm. In the next section, we will show that the convergence rate of the Frank-Wolfe algorithm is exponential when the dictionary is quasi-incoherent and β is large enough so that the expansion x^* is recovered.

B. Rate of convergence

As mentioned in the introduction, in the generic case of Problem (3), the Frank-Wolfe algorithm converges exponentially beyond a certain iteration when the objective function is strongly convex [7] and linearly in the other cases [5]. We prove in Theorem 2, that when the dictionary is quasi-incoherent, the Frank-Wolfe algorithm converges exponentially beyond a certain iteration even though the function we consider is *not* strongly convex.

Theorem 2. *Let Φ be a dictionary, μ its coherence, μ_1 its Babel function, and $y = \Phi x^*$ an m-sparse signal.*
If $m < \frac{1}{2}(\mu^{-1}+1)$ and $\|x^*\|_1 < \beta$, then there exists K such that for all iterations $k \geq K$ of Algorithm 1, we have:

$$\|r_{k+1}\|_2^2 \leq \|r_k\|_2^2(1-\theta)$$

where

$$\theta = \tfrac{1}{16}\left(\tfrac{1-\mu_1(m-1)}{m}\right)\left(1 - \tfrac{\|x^*\|_1}{\beta}\right)^2.$$

Remark 2. *Note that $0 < \theta \leq 1$. Indeed, $m < \frac{1}{2}(\mu^{-1}+1)$ implies $0 < \mu_1(m-1) < 1$ so that $0 < \frac{1}{16m}(1-\mu_1(m-1))\left(1 - \frac{\|x^*\|_1}{\beta}\right)^2 \leq 1$ i.e. $0 < \theta \leq 1$. Thus, Theorem 2 shows exponential convergence.*

Remark 3. *As for Theorem 1, the same result holds when the Exact Recovery Condition (ERC), $\mu_1(m-1) < 1$ and $\|x^*\|_1 < \beta$ hold.*

Proof of Theorem 2. Note that we are in the case where both Theorem 1 and Corollary 1 hold. Let k be an iteration of Algorithm 1. There are two possibles values for γ_k:

a) $\gamma_k = 0$: then $x_k = x_{k+1}$ and subsequently for all $l \geq k$, $x_k = x_l$ and $f(x_k) = f(x_l)$. The convergence of the objective values yields: $f(x_l) = f(x^*) = 0$ for $l \geq k$ and in particular $\|r_{k+1}\|_2^2 = \|r_k\|_2^2 = 0$. Thus Theorem 2 holds.

b) $0 < \gamma_k \leq 1$: by definition of the residual, we have:

$$\begin{aligned} \|r_{k+1}\|_2^2 &= \|y - \Phi x_{k+1}\|_2^2 \\ &= \|r_k - \Phi\gamma_k(s_k - x_k)\|_2^2 \\ &= \|r_k\|_2^2 - 2\gamma_k\langle\Phi(s_k - x_k)), r_k\rangle + \gamma_k^2\|\Phi(s_k - x_k))\|_2^2. \end{aligned}$$

γ_k is the solution of the following optimization problem:

$$\gamma_k = \arg\min_{\gamma \in [0,1]} \|y - \Phi(x_k + \gamma(s_k - x_k))\|_2^2.$$

Notice that $\|\Phi v\|_2 \leq \|v\|_1$ for all vectors v in $\mathbb{R}^n$ since the φ_i are of unit norm ($\|\Phi v\|_2^2 = \sum_{i,j=1}^n v[i]v[j]\langle\varphi_i, \varphi_j\rangle \leq \sum_{i,j=1}^n |v[i]||v[j]| = \|v\|_1^2$). Hence

$$\|y\|_2 = \|\Phi x^*\|_2 \leq \|x^*\|_1 < \beta.$$

We show in Lemma 2 (stated in the appendix) that the solution of the previous problem is then:

$$\gamma_k = \frac{\langle\Phi(s_k - x_k), r_k\rangle}{\|\Phi(s_k - x_k)\|_2^2}.$$

Replacing the value of γ_k in the previous equation we obtain:

$$\|r_{k+1}\|_2^2 = \|r_k\|_2^2 - \frac{\langle\Phi(s_k - x_k), r_k\rangle^2}{\|\Phi(s_k - x_k)\|_2^2}. \tag{4}$$

We shall now lower-bound $\langle\Phi(s_k - x_k), r_k\rangle^2$ and upper-bound $\|\Phi(s_k - x_k)\|_2^2$.

To bound $\|\Phi(s_k - x_k)\|_2^2$, we use $\|\Phi v\|_2 \leq \|v\|_1$ and s_k and x_k are in $\mathcal{B}_1(\beta)$:

$$\|\Phi(s_k - x_k)\|_2^2 \leq \|s_k - x_k\|_1^2 \leq 4\beta^2. \tag{5}$$

To bound $\langle\Phi(s_k - x_k), r_k\rangle^2$, fix $\epsilon = \frac{\beta - \|x^*\|_1}{2} > 0$. As noted in Corollary 1, the iterates x_k converge to x^*, there exists an iteration K such that for every $k \geq K$: $\|x_k - x^*\|_1 \leq \epsilon$. Fix $k \geq K$. Let us define $u \in \mathbb{R}^n$, such that:

$$u[i] = \begin{cases} \frac{\epsilon}{\sqrt{m}\|\Phi^t_{\Lambda_{opt}} r_k\|_2}(\Phi^t r_k)[i] & \text{if } i \in \Lambda_{opt} \\ 0 & \text{otherwise.} \end{cases}$$

One can show that $x_k + u$ is in $\mathcal{B}_1(\beta)$, indeed

$$\begin{aligned} \|x_k + u\|_1 &\leq \|x_k - x^*\|_1 + \|x^*\|_1 + \|u\|_1 \\ &\leq \epsilon + \|x^*\|_1 + \frac{\epsilon}{\sqrt{m}\|\Phi^t_{\Lambda_{opt}} r_k\|_2}\|\Phi^t_{\Lambda_{opt}} r_k\|_1. \end{aligned}$$

Noting that $\|\Phi^t_{\Lambda_{opt}} r_k\|_1 \leq \sqrt{m}\|\Phi^t_{\Lambda_{opt}} r_k\|_2$ (because $\Phi^t_{\Lambda_{opt}} r_k \in \mathbb{R}^{|\Lambda_{opt}|}$ and $|\Lambda_{opt}| \leq m$) leads to:

$$\|x_k + u\|_1 \leq 2\epsilon + \|x^*\|_1 = \beta.$$

We conclude that $x_k + u$ is in $\mathcal{B}_1(\beta)$.

Since $s_k = \arg\min_{s \in \mathcal{B}_1(\beta)} \langle s, \nabla f(x_k)\rangle$ then:

$$\begin{aligned} \langle s_k, \nabla f(x_k)\rangle &\leq \langle x_k + u, \nabla f(x_k)\rangle \\ \langle s_k - x_k, \nabla f(x_k)\rangle &\leq \langle u, \nabla f(x_k)\rangle. \end{aligned}$$

By definition of f: $\nabla f(x_k) = -\Phi^t r_k$, thus:

$$\begin{aligned} \langle s_k - x_k, \Phi^t r_k\rangle &\geq \langle u, \Phi^t r_k\rangle \\ \langle \Phi(s_k - x_k), r_k\rangle &\geq \langle u, \Phi^t r_k\rangle. \end{aligned}$$

Noting that

$$\langle u, \Phi^t r_k\rangle = \sum_{i \in \Lambda_{opt}} \frac{\epsilon}{\sqrt{m}\|\Phi^t_{\Lambda_{opt}} r_k\|_2} (\Phi^t r_k)^2[i] = \frac{\epsilon \|\Phi^t_{\Lambda_{opt}} r_k\|_2}{\sqrt{m}},$$

we conclude:

$$\langle \Phi(s_k - x_k), r_k\rangle \geq \frac{\epsilon}{\sqrt{m}} \|\Phi^t_{\Lambda_{opt}} r_k\|_2.$$

By Theorem 1, r_k is in $\mathrm{span}(\Phi_{\Lambda_{opt}})$ and since the atoms indexed by Λ_{opt} are linearly independent (thus $\lambda^*_{min} > 0$), we obtain:

$$\|\Phi^t_{\Lambda_{opt}} r_k\|_2 \geq \lambda^*_{min} \|r_k\|_2$$

and

$$\langle \Phi(s_k - x_k), r_k\rangle \geq \frac{\epsilon}{\sqrt{m}} \lambda^*_{min} \|r_k\|_2.$$

By Lemma 2.3 of [17], $(\lambda^*_{min})^2 \geq 1 - \mu_1(m-1)$. Since $m < \frac{1}{2}(\mu^{-1} + 1)$ implies $\mu_1(m) + \mu_1(m-1) < 1$ [17], we have $0 < 1 - \mu_1(m-1) < 1$ and deduce that:

$$\langle \Phi(s_k - x_k), r_k\rangle \geq \epsilon \sqrt{\frac{1 - \mu_1(m-1)}{m}} \|r_k\|_2. \tag{6}$$

Plugin in the bounds of Eq. (5) and (6) in Eq. (4), we obtain:

$$\begin{aligned} \|r_{k+1}\|^2 =& \|r_k\|^2 - \frac{\langle \Phi(s_k - x_k), r_k\rangle^2}{\|\Phi(s_k - x_k)\|^2} \\ \|r_{k+1}\|^2 \leq& \|r_k\|^2 \left(1 - \frac{\epsilon^2(1 - \mu_1(m-1))}{4\beta^2 m}\right) \\ \|r_{k+1}\|^2 \leq& \|r_k\|^2 \left(1 - \tfrac{1}{16}\left(\tfrac{1-\mu_1(m-1)}{m}\right)\left(1 - \tfrac{\|x^*\|_1}{\beta}\right)^2\right). \end{aligned}$$

which finishes the proof. □

A natural question that comes from examining Theorem 2 and from the convergence rate of MP and OMP is whether it is possible to guarantee the exponential convergence from the first iteration. The following theorem proves that this is possible if β is large enough.

Theorem 3. *Let Φ be a dictionary of coherence μ and y an m-sparse signal. If $m < \frac{1}{2}(\mu^{-1} + 1)$ and*

$$\beta > 2\|y\|_2 \sqrt{\tfrac{m}{1 - \mu_1(m-1)}},$$

then Algorithm 1 converges exponentially from the first iteration.

We proved in Theorem 2 that when the iterates x_k stay close enough x^* ($\|x_k - x^*\|_1 \leq \epsilon = \frac{\beta - \|x^*\|_1}{2}$), Algorithm 1 converges exponentially. The intuition of Theorem 3 is to choose β large enough so that a similar bound this is guaranteed from the first iteration.

Remark 4. *Let us remark that the assumption $\beta > 2\|y\|_2 \sqrt{\frac{m}{1-\mu_1(m-1)}}$ is stronger than that of Theorem 2 ($\beta > \|x^*\|_1$). Indeed, we have on the one hand: $\|x^*\|_1 \leq \sqrt{m}\|x^*\|_2$ because x^* has non-zero coefficients only in λ_{opt}. On the other hand:*

$$\|y\|_2 = \|\Phi x^*\|_2 \geq \lambda^*_{min} \|x^*\|_2 \geq \sqrt{1 - \mu_1(m-1)} \|x^*\|_2.$$

We conclude

$$\|x^*\|_1 \leq \|y\|_2 \sqrt{\tfrac{m}{1 - \mu_1(m-1)}}. \tag{7}$$

So that $\beta > 2\|y\|_2 \sqrt{\frac{m}{1-\mu_1(m-1)}}$ implies $\beta > \|x^*\|_1$.

Besides, while the assumption of Theorem 2 ($\beta > \|x^*\|_1$) can not be verified beforehand since it depends on the unknown x^*, the assumption of Theorem 3 can be checked since it depends on the dictionary and y.

To prove Theorem 3, the key is to bound uniformly the l_1 norm of the iterates x_k. This is the purpose of the following Lemma.

Lemma 1. *Let Φ be a dictionary, μ its coherence, and $y = \Phi x^*$ an m-sparse signal. If $m < \frac{1}{2}(\mu^{-1} + 1)$, then for each iteration k of Algorithm 1*

$$\|x_k\|_1 \leq 2\|y\|_2 \sqrt{\tfrac{m}{1 - \mu_1(m-1)}}.$$

The proof of Lemma 1 is available in Appendix B. We can now give the proof of Theorem 3.

Proof of Theorem 3. Since $\beta > 2\|y\|_2 \sqrt{\frac{m}{1-\mu_1(m-1)}} \geq \|x^*\|_1$ (see Remark 4), we can re-use the arguments of the proof of Theorem 2. We will do so up to Eq. (5) and only modify the lower bound on $\langle \Phi(s_k - x_k), r_k\rangle^2$.

By definition of s_k,

$$\langle \Phi s_k, r_k\rangle = \beta \max_i |\langle \varphi_i, r_k\rangle|$$

and

$$\begin{aligned} \langle \Phi x_k, r_k\rangle &= \langle x_k, \Phi^t r_k\rangle = \sum_i (x_k)[i](\Phi^t r_k)[i] \\ &\leq \|x_k\|_1 \max_i |(\Phi^t r_k)[i]| \\ &= \|x_k\|_1 \max_i |\langle \varphi_i, r_k\rangle|. \end{aligned}$$

We conclude that:

$$\begin{aligned} \langle \Phi(s_k - x_k), r_k\rangle &= \langle \Phi s_k, r_k\rangle - \langle \Phi x_k, r_k\rangle \\ &\geq \beta \max_i |\langle \varphi_i, r_k\rangle| \left(1 - \frac{\|x_k\|_1}{\beta}\right). \end{aligned}$$

By Lemma 2 of [6]:

$$\max_i |\langle \varphi_i, r_k\rangle| \geq \|r_k\|_2 \sqrt{\tfrac{1 - \mu_1(m-1)}{m}}.$$

Hence:

$$\langle \Phi(s_k - x_k), r_k \rangle \geq \beta \|r_k\|_2 \sqrt{\tfrac{1-\mu_1(m-1)}{m}} \left(1 - \frac{\|x_k\|_1}{\beta}\right). \quad (8)$$

What is left to prove is that $1 - \frac{\|x_k\|_1}{\beta}$ is uniformly bounded away from zero. Using Lemma 1 we obtain:

$$1 - \frac{\|x_k\|_1}{\beta} \geq 1 - \frac{2\|y\|_2}{\beta} \sqrt{\tfrac{m}{1-\mu_1(m-1)}}. \quad (9)$$

By assumption $1 - \frac{2\|y\|_2}{\beta} \sqrt{\frac{m}{1-\mu_1(m-1)}} > 0$, we deduce that for all k:

$$\langle \Phi(s_k - x_k), r_k \rangle \geq \beta \|r_k\|_2 \sqrt{\tfrac{1-\mu_1(m-1)}{m}} (1 - \tau) > 0, \quad (10)$$

with

$$\tau = \tfrac{2\|y\|_2}{\beta} \sqrt{\tfrac{m}{1-\mu_1(m-1)}}. \quad (11)$$

Using this and Eq. (4), we obtain:

$$\begin{aligned} \|r_{k+1}\|_2^2 &= \|r_k\|_2^2 - \frac{\langle \Phi(s_k - x_k), r_k \rangle^2}{\|\Phi(s_k - x_k)\|^2} \\ &\leq \|r_k\|_2^2 - \|r_k\|_2^2 \frac{1 - \mu_1(m-1)}{4m} (1 - \tau)^2. \end{aligned}$$

We conclude that for all k:

$$\|r_{k+1}\|_2^2 \leq \|r_k\|_2^2 \left(1 - \frac{(1 - \mu_1(m-1))}{4m} (1 - \tau)^2\right),$$

with $0 < \left(1 - \frac{(1-\mu_1(m-1))}{4m} (1 - \tau)^2\right) < 1$ which proves the exponential convergence from the first iteration. □

IV. Experiments

Theorem 2 shows that there is exponential convergence when the sparsity m is small enough $m < \frac{1}{2}(\mu^{-1} + 1)$ and β is larger than $\|x^*\|_1$. The goal of this section is to investigate whether these conditions are tight by performing three numerical experiments on synthetic data.

We simulate in Python signals of size $d = 1000$ that are sparse on a dictionary of $n = 2000$ atoms. The dictionary is the union of two orthonormal bases: a DCT-II dictionary [9] and the identity. This dictionary has low coherence ($\mu = 4.5 \times 10^{-2}$), and $m^* = \lceil \frac{1}{2}(\mu^{-1} - 1) \rceil = 11$ is the largest integer such that the condition of Theorem 2 holds. The supports of size m are drawn uniformly at random while the corresponding coefficients are chosen using a standard normal distribution. For each experiment, Algorithm 1 is run for 10000 simulated signals.

The exponential convergence in Theorem 2, is quantified by

$$\|r_k\|_2^2 \leq \|r_{k-1}\|_2^2 (1 - \theta).$$

which is equivalent to

$$\log \|r_k\|_2^2 \leq \log \|y\|_2^2 + k \log(1 - \theta).$$

In the first two experiments, we visualize the convergence rate by displaying the quantity $\log \|r_k\|_2^2$, the convergence being exponential when this is upper-bounded by a line with negative slope (the steepest the slope, the fastest the convergence).

In the first experiment, the values of m and β comply with the conditions of Theorem 2. We fix $\beta = 8\|x^*\|_1$, and $m = m^*$. We draw in Figure 1 the mean and the maximum over the 10000 simulated signals of the function $\log \|r_k\|_2^2$ for each iteration k, and compare it to the theoretical bound in Theorem 2. As expected, the maximum and the mean of the function $\log \|r_k\|_2^2$ can be bounded above by a line with negative slope, and thus converge exponentially. We also notice that in practice, the maximum and the mean are much lower than the theoretical prediction. This suggests that the theoretical bound might be improved in this case.

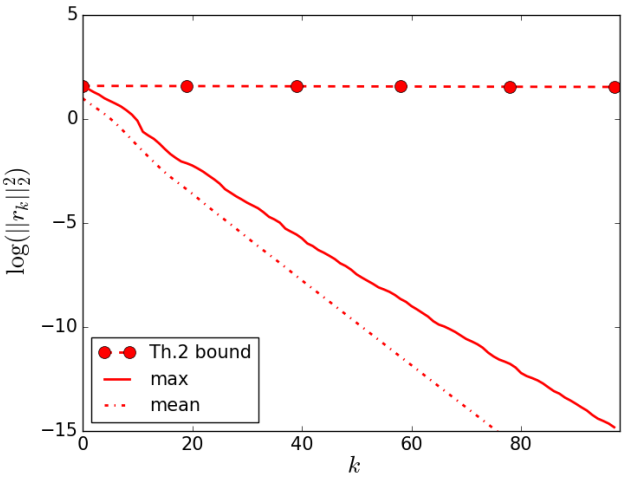

Fig. 1. Evolution of $\log \|r_k\|_2^2$ in Algorithm 1, for $\beta = 8\|x^*\|_1$, and $m = m^* = \lceil \frac{1}{2}(\mu^{-1} - 1) \rceil$. Comparison of the theoretical bound with results obtained on 10000 simulated signals.

In the second experiment, we investigate if the exponential convergence is still possible when the sparsity is larger than $m^* = \lceil \frac{1}{2}(\mu^{-1}+1) \rceil - 1$, i.e., when the condition of Theorem 2 is not satisfied. We fix here $\beta = 8\|x^*\|_1$ and show the maximal value of $\log \|r_k\|_2^2$ for $m = m^*$, $2m^*$, $5m^*$ and $20m^*$ in Figure 2. We observe that exponential convergence still arises at least up to $m = 5m^*$ but probably not for $m = 20m^*$, suggesting that in practice one may reconstruct very fast a larger set of signals than only those being m^*-sparse, and that there might be room for a little improvement in the assumption $m \leq m^*$ in Theorem 2.

In the last experiment, we study the influence of the distance from x^* to $\mathcal{B}_1(\beta)$ on the convergence rate. Indeed Theorem 2 predicts that the convergence slows down when $\|x^*\|_1$ approaches β and does not predict exponential convergence if $\|x^*\|_1 = \beta$.

In this experiment, the sparsity m is fixed to $m = m^*$. We will display the ratio $\frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}$ which if smaller than 1 shows exponential convergence. We show in Figure 3 the mean (over 10000 simulated signals) and theoretical values of the ratio $\frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}$ in two cases: either $\beta = \beta_1 = 1.1\|x^*\|_1$ or $\beta = \beta_2 = 8\|x^*\|_1$. As expected, the ratio is smaller when β is larger. For large β, the ratio stays well below the theoretical bound. This is not the case anymore for β close to $\|x^*\|_1$ suggesting that the theoretical bound may be reached and that the assumption $\beta > \|x^*\|_1$ might be necessary.

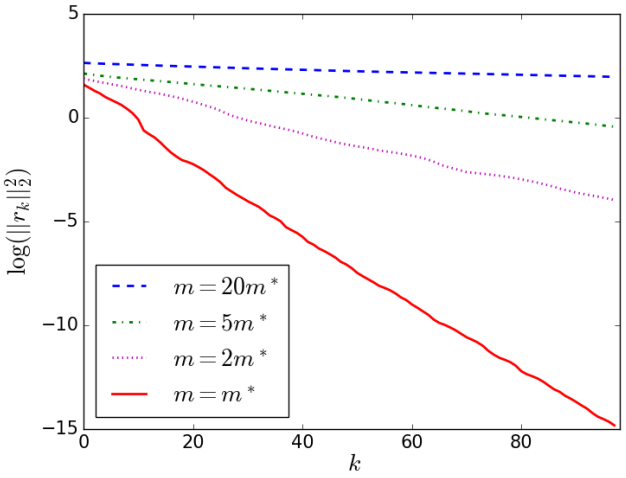

Fig. 2. Evolution the maximum value of $\log \|r_k\|_2^2$ for 10000 simulated signals, with $\beta = 8\|x^*\|_1$ and for different values of m.

Interestingly, one notes here a drop in the mean value for the case $\beta = \beta_2$, at iterations $k \leq m^* = 11$. This drop in the ratio may be explained by the fact that at this point most of the atoms corresponding to the largest expansion coefficients have been selected. But further considerations on this matter go beyond the scope of this paper.

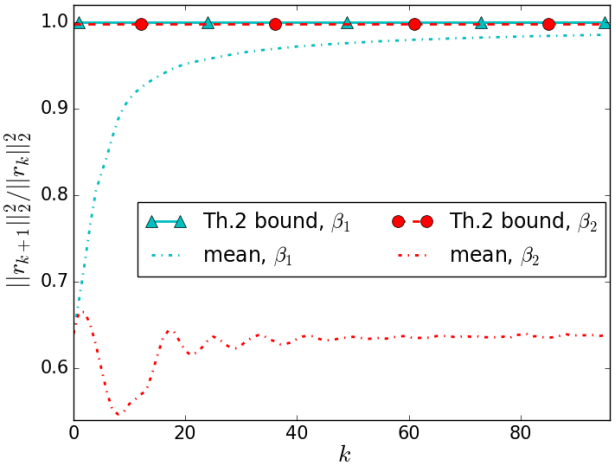

Fig. 3. Evolution of the mean of $\frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}$ for 10000 simulated signals, with $m = m^*$ and for two values of β: $\beta_1 = 1.1\|x^*\|_1$ and $\beta_2 = 8\|x^*\|_1$.

V. Conclusion

We studied in this paper the properties of the Frank-Wolfe algorithm when solving the m-Exact-Sparse problem. We proved that as MP and OMP, when the dictionary is quasi-incoherent, the Frank-Wolfe algorithm picks up only atoms of the support. We also proved that under this same condition, the Frank-Wolfe algorithm converges exponentially. In the experimental part, we have observed the optimality of the obtained bound in terms of the size of the l_1-ball constraining the search space. We have gained some insights on the sparsity bound, suggesting to study its tightness in future work. Extending these results to the case of non-exact-sparse but only compressible signals is also a natural next step.

Appendix A
Proof of Lemma 1

Lemma 1. *Let Φ be a dictionary, μ its coherence, and $y = \Phi x^*$ an m-sparse signal. If $m < \frac{1}{2}(\mu^{-1} + 1)$, then for each iteration k of Algorithm 1*

$$\|x_k\|_1 \leq 2\|y\|_2 \sqrt{\frac{m}{1-\mu_1(m-1)}}.$$

Proof. Indeed, we have on the one hand: $\|x_k\|_1 \leq \sqrt{m}\|x_k\|_2$ because x_k has non-zero coefficients only in λ_{opt} (proved in Theorem 1). On the other hand:

$$\|\Phi x_k\|_2 \geq \lambda^*_{min}\|x_k\|_2 \geq \sqrt{1-\mu_1(m-1)}\|x_k\|_2.$$

We conclude

$$\|x_k\|_1 \leq \sqrt{\frac{m}{1-\mu_1(m-1)}}\|\Phi x_k\|_2.$$

Moreover

$$\begin{aligned}\|\Phi x_k\|_2 &\leq \|\Phi x_k - y\|_2 + \|y\|_2 \\ &\leq \sqrt{2f(x_k)} + \|y\|_2 \\ &\leq \sqrt{2f(x_0)} + \|y\|_2 \\ &\leq \|y\|_2 + \|y\|_2 \\ &\leq 2\|y\|_2,\end{aligned}$$

where the third line holds because by construction of the Frank-Wolfe algorithm, the sequence $\{f(x_k)\}_k$ is non increasing. So we conclude that $\|x_k\|_1 \leq 2\|y\|_2 \sqrt{\frac{m}{1-\mu_1(m-1)}}$. □

Appendix B
Proof of Lemma 2

Lemma 2. *For any iteration k of Algorithm 1, if $\gamma_k \neq 0$, and if $\|y\|_2 < \beta$ then:*

$$\gamma_k = \frac{\langle \Phi(s_k - x_k), r_k \rangle}{\|\Phi(s_k - x_k)\|_2^2}.$$

Proof. Recall that

$$\gamma_k = \arg\min_{\gamma \in [0,1]} \|y - \Phi(x_k + \gamma(s_k - x_k))\|_2^2$$

and define

$$\gamma_k^* = \arg\min_{\gamma \in \mathbb{R}} \|y - \Phi(x_k + \gamma(s_k - x_k)\|_2^2.$$

Note that

$$\gamma_k^* = \frac{\langle \Phi(s_k - x_k), r_k \rangle}{\|\Phi(s_k - x_k)\|_2^2},$$

so we wish to prove that $\gamma_k = \gamma_k^*$.

Because γ_k is the solution of the same minimization problem as γ_k^* but restricted on the interval $[0, 1]$, we have only three possibilities: (i) $\gamma_k^* \geq 1$ and $\gamma_k = 1$, (ii) $0 < \gamma_k = \gamma_k^* < 1$, (iii) $\gamma_k^* \leq 0$ and $\gamma_k = 0$. Here we assume that $\gamma_k \neq 0$ so the last possibility (iii) is ruled out. What is left to do to finish the proof is to rule out the first possibility: (i) $\gamma_k^* \geq 1$ and $\gamma_k = 1$.

To do so, consider these two different cases:

- $k = 0$: since $x_0 = 0$ and $r_0 = y$,

$$\gamma_0^* = \frac{\langle \Phi s_0, y \rangle}{\|\Phi s_0\|_2^2} \leq \frac{\|y\|_2}{\beta}.$$

  Since $\|y\|_2 < \beta$, we have $\gamma_0^* < 1$. Moreover, by construction of s_0, $\langle \Phi s_0, y \rangle > 0$ so $0 < \gamma_0^* < 1$. We conclude we are in case (ii) and $\gamma_0 = \gamma_0^*$.
- $k \neq 0$: assume that $\gamma_k = 1$. We then have $x_{k+1} = s_k$. By construction of the Frank-Wolfe algorithm we have:

$$f(x_{k+1}) = f(s_k) \leq f(x_k) \leq \cdots \leq f(x_1) = f(\gamma_0 s_0).$$

  Since we proved that $\gamma_0 \neq 1$, we have:

$$f(\gamma_0 s_0) < f(s_0).$$

  This implies $f(s_k) < f(s_0)$, that is:

$$\|y - \Phi s_k\|_2^2 < \|y - \Phi s_0\|_2^2$$
$$\langle \Phi s_0, y \rangle < \langle \Phi s_k, y \rangle.$$

  Since $s_0 = \operatorname{sign}(\langle \varphi_{i_0}, y \rangle) \beta e_{i_0}$, both sides of the previous equation are positive:

$$0 < \langle \Phi s_0, y \rangle < \langle \Phi s_k, y \rangle$$

  This is clearly a contradiction because $s_0 = \arg\max_{s \in \mathcal{B}_1(\beta)} \langle \Phi s, y \rangle$. We conclude that $0 < \gamma_k < 1$ so that we are again in case (ii) where $\gamma_k = \gamma_k^*$.

We conclude that if $\gamma_k > 0$ and $\|y\|_2 < \beta$ then $\gamma_k = \frac{\langle \Phi(s_k - x_k), r_k \rangle}{\|\Phi(s_k - x_k)\|_2^2}$. □

ACKNOWLEDGMENT

The authors would like to thank the Agence Nationale de la Recherche under grant JCJC MAD (ANR-14-CE27-0002) which supported this work.

REFERENCES

[1] Thomas Blumensath and Mike E Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.
[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
[3] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.
[4] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.
[5] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.
[6] Rémi Gribonval and Pierre Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, 52(1):255–261, 2006.
[7] Jacques Guélat and Patrice Marcotte. Some comments on wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.
[8] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. In *Artificial Intelligence and Statistics*, pages 860–868, 2017.
[9] John Makhoul. A fast cosine transform in one and two dimensions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):27–34, 1980.
[10] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.
[11] Alan Miller. *Subset selection in regression*. Chapman and Hall/CRC, 2002.
[12] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44, 1993.
[13] Bhaskar D Rao and Kenneth Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Transactions on signal processing*, 47(1):187–200, 1999.
[14] Vladimir Temlyakov. *Greedy approximation*, volume 20. Cambridge University Press, 2011.
[15] Vladimir N Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3(1), 2003.
[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
[17] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.
[18] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.